Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

R.R.P. van Nooyen

An exponential fitting method in two dimensions

Department of Numerical Mathematics Report NM-R9005 February

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Research (N.W.O.).

An Exponential Fitting Method in Two Dimensions

R. R. P. van Nooyen

CWI, Centre for Mathematics and Computer Science,
P. O. box 4079, 1009 AB Amsterdam, The Netherlands

We discuss a Petrov-Galerkin form of exponential fitting for a mixed finite element formulation of the semiconductor continuity equations on a rectangular domain. We give conditions, that ensure $\mathcal{O}(h)$ bounds for the $L^2(\Omega)$ norm of the error. The constants in the error bound do not contain exponentials of the electrical potential.

1980 AMS Subject Classification: 65N30
1987 CR category: G. 1.8
Key Words and Phrases: Mixed Finite Elements, Semiconductor Continuity Equation, Exponential Fitting.
Note: This report will be submitted for publication elsewhere.

1 Introduction.

The use of a form of exponential fitting for the semiconductor continuity equations is suggested by the success of the Scharfetter-Gummel discretisation[1] in one dimension and variations on that discretisation in two dimensions. Numerous derivations of Scharfetter-Gummel type discretisations are given in the literature, for instance by Selberherr[2], Markowich[3], Bank et al.[4], Brezzi et al.[5], and others. This paper extends a one dimensional exponential fitting technique, discussed by Hemker[6], to the two dimensional problem.

In section 2 we introduce the version of the semiconductor continuity equation, that we use in this paper. We introduce several bilinear forms, related with the coefficients in this equation. In section 3 we prepare the way for the error estimates. The next section discusses the discretisation. In section 5 we collect some technical results and in section 6 we derive two error estimates. These error estimates are based on the techniques used by Douglas and Roberts[7]. The proofs in section 6 use all properties of our special discrete system, in particular the use of a quadrature rule for the approximation of certain integrals in the discrete system. To show the method in action, in the next section we use it to find a discretisation for a simple one dimensional equation, where we can compare our result with a known exact solution. In the last section we discuss our findings.

2 The equation.

We consider the following problem, find $u \in H^2(\Omega)$ such that:

$$-div\,(\frac{1}{\alpha}(grad\ u + u\vec{\beta})) + \gamma u = f \quad \text{on} \quad \Omega \ \text{ and} \tag{1}$$

$$u = -g \ \text{ on } \ \partial\Omega\ .$$

Where Ω is a bounded rectangular domain in $\mathbb{R}^2$. We impose the following restrictions on the coefficients:

$$\alpha \in W^{1,\infty}(\Omega) \ \text{ and } \ \exists\ A \in \mathbb{R} : \alpha \geqslant A > 0 \ \text{ on } \ \Omega\ , \tag{2}$$

$$\frac{1}{\alpha} \in W^{1,\infty}(\Omega) \ \text{ on } \ \Omega\ , \tag{3}$$

$$\vec{\beta} = (\beta_1,\beta_2)^T \text{ with } \beta_1,\beta_2 \in W^{1,\infty}(\Omega)\ , \tag{4}$$

$$\gamma \in W^{1,\infty}(\Omega) \text{ and } \gamma \geqslant 0 \text{ on } \Omega . \tag{5}$$

Where $W^{1,\infty}(\Omega)$, $H^2(\Omega) = W^{1,2}(\Omega)$ are the usual Sobolev spaces[8], and

$$H(div;\Omega) := \{ \tau \in L^2(\Omega)^2 \mid div\, \tau \in L^2(\Omega) \} ,$$

with scalar product

$$(\sigma,\tau)_{H(div;\Omega)} = \int_\Omega \sigma\cdot\tau\, d\mu + \int_\Omega div\, \sigma\, div\, \tau\, d\mu ,$$

is a Hilbert space (see also Girault and Raviart, [9] formula 2.15 in section 2.2). We assume, that the equation has a solution, that $f \in L^2(\Omega)$, $g \in L^2(\partial\Omega)$ and, that

$$\int_{\partial\Omega} g\tau\cdot\vec{n}_{\partial\Omega}\, d(\partial\Omega) \leqslant C \| \tau \|_{H(div;\Omega)} \quad \forall \ \tau \in H(div;\Omega) .$$

If the Einstein relations[2] hold, then the stationary semiconductor continuity equations take the form (1). Here $\vec{\beta}$ corresponds with the gradient of the electric potential, the term γu corresponds to a linear approximation to the recombination term and $1/\alpha$ corresponds to the electron or hole mobility. The exact correspondence depends on the choice of scaling[10].

We use the following bilinear forms to formulate the weak mixed form of this equation,

$$(s,t) = \int_\Omega s\, t\, d\mu \quad \forall \ s,t \in L^2(\Omega) \ ,$$

$$a(\sigma,\tau) = \int_\Omega \alpha\sigma\cdot\tau\, d\mu \quad \forall \ \sigma,\tau \in H(div;\Omega) ,$$

$$b(\sigma,t) = \int_\Omega \vec{\beta}\cdot\sigma\, t\, d\mu \quad \forall \ \sigma \in H(div;\Omega) , t \in L^2(\Omega) \ ,$$

$$c(s,t) = \int_\Omega \gamma\, s\, t\, d\mu \quad \forall \ s,t \in L^2(\Omega) \ ,$$

$$< g,h > \ = \int_{\partial\Omega} g\, h\, d\mu \quad \forall \ g,h \in L^2(\partial\Omega) .$$

Given these definitions, we see immediately, that any solution $u \in H^2(\Omega)$ of (1) generates a solution $(\sigma,u) \in H(div;\Omega) \times L^2(\Omega)$ of

$$a(\sigma,\tau) - (div\, \tau,u) + b(\tau,u) = \ < g,\tau\cdot\vec{n}_{\partial\Omega} > \quad \forall \ \tau \in H(div;\Omega) , \tag{6a}$$

$$(div\, \sigma,t) + c(u,t) = (f,t) \quad \forall \ t \in L^2(\Omega) \ . \tag{6b}$$

Where $\sigma = -\frac{1}{\alpha}(grad\, u + u\vec{\beta})$.

To simplify the notation, we denote the Cartesian product of a normed linear space E with itself by E in bold faced type, $\mathbf{E} := E \times E$. We define

$$\| (\mu_1,\mu_2)^T \|_{\mathbf{E}} := (\sum_{i=1}^{2} \| \mu_i \|_E^2)^{1/2} \quad \forall \ (\mu_1,\mu_2)^T \in \mathbf{E} .$$

3 Preparations.

In this section, we introduce a partition of our domain and we define the adjoint problem of (1), which we use in the derivation of one of our error estimates. Next, we introduce several special projections, that are needed in the definition of our approximation spaces and in the derivation of the error estimates. Finally we give a general error estimate for mappings, that leave all polynomials of degree lower than $m+1$ invariant.

3.1. The partitioning of the domain.

Before we treat our discretisation, we define our approximation space. We introduce a set of rectangular, open subdomains Ω_k of Ω, where $k \in K$, K an index set. We assume, that the subdivision is regular and has the following two properties,

$$\bigcup_{k \in K} \bar{\Omega}_k = \bar{\Omega} \text{ and } \quad \forall\ k,l \in K, k \neq l : \Omega_k \cap \Omega_l = \emptyset \ .$$

On Ω, we use orthonormal coordinates, with the unit vectors $\vec{e}_1$ and $\vec{e}_2$ parallel to the edges of Ω. We define $\tau_i := \tau \cdot \vec{e}_i$ for $\tau \in \mathbf{L}^2(\Omega)$ and $x_i := \vec{x} \cdot \vec{e}_i$ for $\vec{x} \in \mathbb{R}^2$. If we define $\vec{x}_k$ to be the lower left corner of Ω_k and $\vec{x}_k + \vec{h}_k$ the upper right corner of Ω_k, then

$$\Omega_k = \{ \vec{x} \in \mathbb{R}^2 \mid 0 < (\vec{x} - \vec{x}_k) \cdot \vec{e}_i < \vec{h}_k \cdot \vec{e}_i \quad \text{for} \quad i = 1,2 \} \ . \tag{7}$$

We use the notation χ_k for the characteristic function of Ω_k. (The characteristic function of a set is the function that is equal to one in all points of the set and zero elsewhere). The edges of Ω_k are the sets:

$$\Gamma_{k,i,j} = \{ \vec{x} \in \bar{\Omega}_k \mid \vec{x} \cdot \vec{e}_i = (\vec{x}_k + j\vec{h}_k) \cdot \vec{e}_i \} \quad \text{for} \quad i = 1,2 \ , j = 0,1 \ . \tag{8}$$

$\chi_{k,i,j}$ is the characteristic function of side $\Gamma_{k,i,j}$.

3.2. The adjoint problem.

We use the following definition for the adjoint problem of (1) (cf. Douglas and Roberts [7]),

$$w \in \mathrm{H}^2(\Omega) \ , \tag{9}$$

$$-div\,(\frac{1}{\alpha} grad\ w) + \frac{\vec{\beta}}{\alpha} \cdot grad\ w + \gamma w = f \ ,$$

$$w = 0 \text{ on } \partial\Omega \ .$$

The adjoint problem is called regular, if there is a unique solution w for every $f \in \mathrm{L}^2(\Omega)$ and this solution satisfies $\| w \|_{\mathrm{H}^2(\Omega)} \leqslant C \| f \|_{\mathrm{L}^2(\Omega)}$ for every $f \in \mathrm{L}^2(\Omega)$.

Remark: Both in the above equation and in the rest of this report, the upper case C, without a subscript, denotes a generic constant. It may have a different value at each appearance.

The weak mixed form of the adjoint problem is:

$$(\tau, w) \in \mathrm{H}(div;\Omega) \times \mathrm{L}^2(\Omega) \ , \tag{10}$$

$$a(\tau,\sigma) - (div\ \sigma, w) = 0 \quad \forall \ \sigma \in \mathrm{H}(div;\Omega) \text{ and} \tag{10a}$$

$$(div\ \tau, t) - b(\tau, t) + c(w,t) = (f,t) \quad \forall \ t \in \mathrm{L}^2(\Omega) \ . \tag{10b}$$

Any solution $w \in \mathrm{H}^2(\Omega)$ of (9) will generate a solution $(-\frac{1}{\alpha} grad\ w, w)$ of this problem. If (9) is regular, then this solution satisfies $\| w \|_{\mathrm{H}^2(\Omega)} + \| \tau \|_{\mathrm{H}^1(\Omega)} \leqslant C \| f \|_{\mathrm{L}^2(\Omega)}$.

3.3. Some projections.

We introduce several local projections, we use these to define four mappings, P_h, $\mathbf{P}_h$, Π_h and $\tilde{\Pi}_h$. First, we define $P[\Omega_k]$ to be the orthogonal projection from $\mathrm{L}^2(\Omega_k)$ to the space of constant functions on Ω_k, and we define $P[\Gamma_{k,i,j}]$ to be the orthogonal projection from $\mathrm{L}^2(\Gamma_{k,i,j})$ to the space of constant functions on $\Gamma_{k,i,j}$.

We use $P[\Omega_k]$ to create two global mappings, $P_h \colon \mathrm{L}^2(\Omega) \to \mathrm{L}^2(\Omega)$,

$$P_h f = \sum_{k \in K} \chi_k P[\Omega_k](f) \quad \forall \ f \in \mathrm{L}^2(\Omega) \ , \tag{11a}$$

and $\mathbf{P}_h \colon \mathbf{L}^2(\Omega) \to \mathbf{L}^2(\Omega)$,

$$\mathbf{P_h}\vec{\beta} = \sum_{k \in K} \chi_k \left[P[\Omega_k](\vec{\beta} \cdot \vec{e}_1)\vec{e}_1 + P[\Omega_k](\vec{\beta} \cdot \vec{e}_2)\vec{e}_2 \right] \quad \forall \ \vec{\beta} \in \mathbf{L}^2(\Omega) \ . \tag{11b}$$

Next, we introduce two mappings, based on $P[\Gamma_{k,i,j}]$. These mappings have as their domain the space Σ,

$$\Sigma := \{ \tau \in \mathrm{H}(div;\Omega) \mid \tau|_{\partial\Omega_k}\cdot\vec{n}_{\partial\Omega_k} \in \mathrm{L}^2(\partial\Omega_k) \quad \forall\ k \in K \} .$$

This space is similar to the one introduced by Roberts and Thomas in formula (1.10) of their report[11].

To simplify the definition of these mappings, we introduce local coordinates on each cell Ω_k,

$$\vec{\xi} := \begin{bmatrix} (x_1 - x_{k,1})/h_{k,1} \\ (x_2 - x_{k,2})/h_{k,2} \end{bmatrix} . \tag{12}$$

The mappings are defined as follows:

$$\Pi_h\tau = \sum_{k \in K} \chi_k \sum_{i=1}^{2} \Big[(1-\xi_{k,i})P[\Gamma_{k,i,0}](\tau_i) + \xi_{k,i}P[\Gamma_{k,i,1}](\tau_i)\Big]\vec{e}_i , \tag{13}$$

$$\tilde{\Pi}_h\tau = \sum_{k \in K} \chi_k \sum_{i=1}^{2} \Big[(1-\eta_{k,i})\ P[\Gamma_{k,i,0}](\tau_i) + \eta_{k,i}P[\Gamma_{k,i,1}](\tau_i)\Big]\vec{e}_i , \tag{14}$$

where $i=1,2$ denotes the horizontal or vertical direction respectively, and where

$$\eta_{k,i} = \begin{cases} \dfrac{\exp(\xi_{k,i}h_{k,i}P[\Omega_k](\beta_i)) - 1}{\exp(h_{k,i}\ P[\Omega_k](\beta_i)) - 1} & \mathit{if}\ P[\Omega_k](\beta_i) \neq 0 , \\ \xi_{k,i} & \mathit{if}\ P[\Omega_k](\beta_i) = 0 . \end{cases}$$

So, for $\Pi_h\tau$ we get the i^{th} component on Ω_k, by linear interpolation between the projections of this component on the two sides orthogonal to $\vec{e}_i$. For $\tilde{\Pi}_h\tau$ however, we obtain the same component by using an exponential function to interpolate between the projections of this component on the two sides orthogonal to $\vec{e}_i$.

We introduce the following finite dimensional function spaces,

$$V_h = \Pi_h(\Sigma)\ , \ W_h = P_h(\mathrm{L}^2(\Omega)) \ \text{and}\ X_h = \tilde{\Pi}_h(\Sigma) .$$

$V_h \times W_h$ is the Raviart-Thomas-Nedelec space of index 0 for rectangles. This space was described by Douglas and Roberts, [7] Raviart and Thomas[12] and, for $\Omega \subset \mathbb{R}^3$, by Nedelec[13]. In this paper we use the usual space, $V_h \times W_h$, as the trial function space and $X_h \times W_h$ as the test function space. In effect, we use exponential test functions instead of the usual linear test functions. Thus, we obtain a kind of Petrov-Galerkin discretisation for mixed finite elements.

3.4. Error estimates for projections.

Later we need the classical projection estimates from Ciarlet and Raviart[14]. We take $1 \leqslant p \leqslant \infty$, $0 \leqslant l \leqslant m+1$ and $G \in \mathscr{L}(\mathrm{W}^{m+1,p}(\Omega_k),\mathrm{W}^{l,p}(\Omega_k))$. We assume, that G leaves polynomials of degree less than $m+1$ invariant. We must distinguish two cases, if $p=\infty$, then the estimates given below always hold, but, if $1 \leqslant p < \infty$ then these estimates only hold for $|\vec{h}_k|$ small enough.

$$\forall\ u \in \mathrm{W}^{\mathrm{m+1,p}}(\Omega_\mathrm{k}) : \|u - Gu\|_{\mathrm{W}^{l,p}(\Omega_k)} \leqslant C(l,m,p,k)\|u\|_{\mathrm{W}^{m+1,p}(\Omega_k)}\frac{h^{m+1}_{\max,k}}{h^{l}_{\min,k}} , \tag{15}$$

where $h_{k,\max} = \max(|h_{k,1}|,|h_{k,2}|)$, $h_{k,\min} = \min(|h_{k,1}|,|h_{k,2}|)$. For a finite index set K and $u \in \mathrm{W}^{\mathrm{m+1,p}}(\Omega)$, this implies:

$$\|u - Gu\|_{\mathrm{W}^{l,p}(\Omega)} \leqslant C(l,m,p)\|u\|_{\mathrm{W}^{m+1,p}(\Omega)}\frac{h^{m+1}_{\max}}{h^{l}_{\min}} , \tag{16}$$

where $h_{\max} = \max\limits_{k \in K} h_{k,\max}$, $h_{\min} = \min\limits_{k \in K} h_{k,\min}$. This corresponds to theorem 5 of Ciarlet and

Raviart[14].

4 The discretisation.

We replace the coefficients α, $\vec{\beta}$ and γ by two dimensional step functions. To write our modified problem in weak form, we need to define three new bilinear forms,

$$\bar{a}(\sigma,\tau) = \int_{\Omega_k} \sigma{\cdot}\tau P_h\alpha \, d\mu \quad \forall \ \sigma,\tau \in \Sigma \, ,$$

$$\bar{b}(\sigma,t) := \int_{\Omega} t\sigma{\cdot}\mathbf{P}_h\vec{\beta} \, d\mu \quad \forall \ \sigma \in \Sigma, t \in \mathrm{L}^2(\Omega) \ ,$$

$$\bar{c}(\sigma,t) := \int_{\Omega} stP_h\gamma \, d\mu \quad \forall \ s,t \in \mathrm{L}^2(\Omega) \ .$$

Then we replace $\bar{a}$ by $\bar{a}_q$, an approximation that takes into account the quadrature, where

$$\bar{a}_q(\sigma,\tau) = \sum_{k \in K} P[\Omega_k](\alpha) \frac{\mu(\Omega_k)}{2} \sum_{i=1}^{2} \left[P[\Gamma_{k,i,0}](\sigma_i\tau_i) + P[\Gamma_{k,i,1}](\sigma_i\tau_i) \right] \quad \forall \ \sigma \, , \tau \in \Sigma \, ; \tag{17}$$

$\bar{a}_q$ determines an approximation to $\bar{a}$ by replacing the integral of $\sigma_i \ \tau_i$ over Ω_k by the average of the projections of this expression on the two sides orthogonal to $\vec{e}_i$.

We approximate the solution (σ,u) of (6) by $(\sigma_h,u_h) \in V_h \times W_h$, where

$$\bar{a}_q(\sigma_h,\tau) - (u_h, div \ \tau) + \bar{b}(\tau,u_h) = \ < \tau{\cdot}\vec{n}_{\partial\Omega}, g > \quad \forall \ \tau \in X_h \, , \tag{18a}$$

$$(div \ \sigma_h,t) + \bar{c}(u_h,t) = (f,t) \quad \forall \ t \in W_h \ . \tag{18b}$$

If we use $\bar{a}$ in stead of $\bar{a}_q$, then our discrete problem does not always yield an M-matrix. Consider, for instance, the corresponding discretisation on a uniform mesh with mesh-width h in one dimension with $\alpha = 1$, $\vec{\beta} = \vec{0}$ and γ constant. If $\alpha\gamma h^2/6 > 1$, then the off-diagonal elements of the discretisation matrix for u_h after elimination of σ_h through static condensation have the same sign as the elements on the diagonal.

The idea of using linear trial functions and exponential test functions was used by Hemker for singularly perturbed two point boundary problems[6]. For the one dimensional case, the introduction of exponential test functions follows from the requirement, that the Green's function of the problem can be approximated by the test functions.

In the following sections, we prove, that the solution of our discretisation (18) is an $\mathcal{O}(h)$ approximation to the solution of our original problem.

5 Several technical results.

This section contains some technical results, collected for later reference.

5.1. The properties of Π_h and $\tilde{\Pi}_h$.

We start of with three lemmas for the two mappings Π_h and $\tilde{\Pi}_h$ from Σ into itself.

Lemma 1.

$$\tilde{\Pi}_h \circ \Pi_h = \tilde{\Pi}_h \, , \tag{19a}$$

$$\Pi_h \circ \tilde{\Pi}_h = \Pi_h \, , \tag{19b}$$

$$(div \ \sigma, P_h t) = (div \ \Pi_h\sigma, t) \quad \forall \ \sigma \in \Sigma \, , t \in \mathrm{L}^2(\Omega) \ , \tag{19c}$$

$$< \Pi_h\tau{\cdot}\vec{n}_{\partial\Omega} > \ = \ < \tilde{\Pi}_h\tau{\cdot}\vec{n}_{\partial\Omega} > \quad \forall \ \tau \in \Sigma \, . \tag{19d}$$

Proof.
Both mappings are based on the same projections $P[\Gamma_{k,i,j}]$, so (19a) and (19b) are trivial.

To prove (19c) we use a special case of Green's theorem:

$$\int_{\Omega_k} div\ \sigma\ d\mu = \sum_{i=1}^{2} \frac{\mu(\Omega_k)}{h_{k,i}} \left[P[\Gamma_{k,i,1}](\sigma_i) - P[\Gamma_{k,i,0}](\sigma_i)\right] .$$

If we combine this with the definition of Π_h, the proof of (19c) is complete. (19d) follows immediately from the definitions. □

In lemma 2, we give estimates for the norm of the image of $\sigma \in \Sigma$ under Π_h and $\tilde{\Pi}_h$.

Lemma 2.
If $\sigma \in \Sigma$ and we define $a_{k,i} = P[\Gamma_{k,i,0}](\sigma_i)$ and $b_{k,i} = P[\Gamma_{k,i,1}](\sigma_i)$, then the following inequalities hold for $\| \Pi_h \sigma \|_{L^2(\Omega_k)}$ and $\| \tilde{\Pi}_h \sigma \|_{L^2(\Omega_k)}$,

$$\frac{\mu(\Omega_k)}{6} \sum_{i=1}^{2} (a_{k,i}^2 + b_{k,i}^2) \leq \| \Pi_h \sigma \|^2_{L^2(\Omega_k)} \leq \frac{\mu(\Omega_k)}{2} \sum_{i=1}^{2} (a_{k,i}^2 + b_{k,i}^2) . \tag{20a}$$

$$\| \tilde{\Pi}_h \sigma \|^2_{L^2(\Omega_k)} \leq 2\mu(\Omega_k) \sum_{i=1}^{2} (a_{k,i}^2 + b_{k,i}^2) \leq 12 \| \Pi \sigma \|^2_{L^2(\Omega_k)} . \tag{20b}$$

Proof.
Formula (20a) follows immediately from

$$\Pi_h \sigma |_{\Omega_k} = \sum_{i=1}^{2} \left[(1-\xi_{k,i})a_{k,i} + \xi_{k,i} b_{k,i}\right] \vec{e}_i .$$

By definition,

$$\tilde{\Pi} \sigma |_{\Omega_k} = \sum_{i=1}^{2} \left[(1-\eta_{k,i})a_{k,i} + \eta_{k,i} a_{k,i}\right] \vec{e}_i ,$$

so

$$\int_{\Omega_k} \tilde{\Pi} S \cdot \tilde{\Pi} S\ d\mu = \sum_{i=1}^{2} \left[\int_{\Omega_k} (a_{k,i}^2 (1-\eta_{k,i})^2 + b_{k,i}^2 \eta_{k,i}^2 + 2a_{k,i} b_{k,i} (1-\eta_{k,i}) \eta_{k,i})\right] \leq$$

$$2\mu(\Omega_k) \sum_{i=1}^{2} (a_{k,i}^2 + b_{k,i}^2) .$$

This is the first inequality of (20b). The second inequality follows (20a). □

Lemma 3 gives an estimate for $\| (\tilde{\Pi}_h - \Pi_h) \Pi_h \sigma \|_{L^2(\Omega_k)}$.

Lemma 3.
If $\sigma \in \Sigma$ and we define $a_{k,i} = P[\Gamma_{k,i,0}](\sigma_i)$ and $b_{k,i} = P[\Gamma_{k,i,1}](\sigma_i)$, then

$$\| (\tilde{\Pi}_h - \Pi_h) \Pi_h \sigma \|_{L^2(\Omega_k)} \leq \tag{21}$$

$$\min\left[2\sqrt{2}\ , 2\sqrt{6} \max(h_{k,1} | P[\Omega_k](\beta_1) |\ , h_{k,2} | P[\Omega_k](\beta_2) |) \right] \| \Pi_h \sigma \|_{L^2(\Omega_k)} .$$

Proof.

$$\| (\Pi_h - \tilde{\Pi}_h) \Pi_h \sigma \|^2_{L^2(\Omega_k)} =$$

$$\int_{\Omega_k} \sum_{i=1}^{2} \left[(\eta_{k,i} - \xi_{k,i}) (a_{k,i} - b_{k,i})\right]^2 d\mu \leq$$

$$\int_{\Omega_k} \left[\sum_{i=1}^{2} (\eta_{k,i} - \xi_{k,i})^2\right] \left[\sum_{i=1}^{2} (a_{k,i} - b_{k,i})^2\right] d\mu \leq$$

$$2 \int_{\Omega_k} \left[\sum_{i=1}^{2} (\eta_{k,i} - \xi_{k,i})^2\right] d\mu \left[\sum_{i=1}^{2} (a_{k,i}^2 + b_{k,i}^2)\right] .$$

This implies,

$$\|(\Pi_h-\tilde{\Pi}_h)\Pi_h\sigma\|^2_{\mathbf{L}^2(\Omega_k)} \leqslant \frac{4\mu(\Omega_k)}{3}\left[\sum_{i=1}^{2}(a^2_{k,i}+b^2_{k,i})\right] \leqslant$$
$$8\,\|\Pi_h\sigma\|^2_{\mathbf{L}^2(\Omega_k)}\ .$$

Furthermore,

$$\int_{\Omega_k}\sum_{i=1}^{2}(\eta_{k,i}-\xi_{k,i})^2\,d\mu =$$
$$\int_{\Omega_k}\sum_{i=1}^{2}\left[\frac{h_{k,i}(\exp(P[\Omega_k](\beta_i)x_i)\,-\,1)\,-\,x_i(\exp(P[\Omega_k](\beta_i)h_{k,i})\,-\,1)}{h_{k,i}(\exp(P[\Omega_k](\beta_i)h_{k,i})\,-\,1)}\right]^2 d\mu\ .$$

To obtain the desired result, we need the following general inequality,

$$\left[\frac{x(\exp(bh)-1)\,-\,h(\exp(bx)-1)}{h(\exp(bh)-1)}\right]^2 \leqslant (hb)^2\ ,$$

where $x,h,b\in\mathbb{R}$, and $0\ <\ x\ <\ h$. We derive this inequality,

$$|x(\exp(bh)-1)\,-\,h(\exp(bx)-1)|\ =\ |\int_{w=0}^{x}\int_{z=0}^{h} b\exp(bz)\,-\,b\exp(bw)\,dz\,dw\,|\ =$$
$$|\int_{w=0}^{x}\int_{z=0}^{h}\int_{v=w}^{z} b^2\exp(bv)\,dv\,dz\,dw\,|\ \leqslant\ \int_{w=0}^{x}\int_{z=0}^{h}|\int_{v=w}^{z} b^2\exp(bv)\,dv\ |\,dz\,dw\ \leqslant$$
$$\int_{w=0}^{h}\int_{z=0}^{h}|\int_{v=w}^{z} b^2\exp(bv)\,dv\ |\,dz\,dw\ \leqslant\ \int_{w=0}^{h}\int_{z=0}^{h}\int_{v=0}^{h} b^2\exp(bv)\,dv\,dz\,dw\ \leqslant$$
$$h^2b(\exp(bh)-1)\ .$$

Hence

$$\|(\Pi_h-\tilde{\Pi}_h)\Pi_h\sigma\|^2_{\mathbf{L}^2(\Omega_k)} \leqslant 2\mu(\Omega_k)\left[\sum_{i=1}^{2}(h_{k,i}P[\Omega_k](\beta_i))^2\right]\left[\sum_{i=1}^{2}(a^2_{k,i}+b^2_{k,i})\right] \leqslant$$
$$12\left[\sum_{i=1}^{2}(h_{k,i}P[\Omega_k](\beta_i))^2\right]\|\Pi_h\sigma\|^2_{\mathbf{L}^2(\Omega_k)}\ .$$

This completes our proof. □

5.2. The properties of a.

We mention some obvious properties of a. If $\sigma,\tau\in\mathbf{L}^2(\Omega)$, then the restrictions for α, as given in (2), imply, that

$$a(\sigma,\tau)\ \leqslant\ \|\alpha\|_{\mathrm{L}^\infty(\Omega)}\,\|\sigma\|_{\mathbf{L}^2(\Omega)}\,\|\tau\|_{\mathbf{L}^2(\Omega)}\ ,$$
$$A\,\|\tau\|^2_{\mathbf{L}^2(\Omega)}\ \leqslant\ a(\tau,\tau)\quad\forall\ \ \tau\in\mathbf{L}^2(\Omega)\ .$$

We see, that a is both $\mathbf{L}^2(\Omega)$-bounded and $\mathbf{L}^2(\Omega)$-elliptic. In the next two sections we discuss the corresponding properties for $\bar{a}$ and $\bar{a}_q$.

5.3. The properties of $\bar{a}$.

Lemma 4 shows, that $\bar{a}$ is $\mathbf{L}^2(\Omega)$-bounded and $\mathbf{L}^2(\Omega)$-elliptic.

Lemma 4.
If $\alpha\in\mathrm{W}^{1,\infty}(\Omega)$, $\alpha\ \geqslant\ A\ >\ 0$ on Ω, $\bar{a}(\sigma,\tau):=\int_{\Omega}P_h(\alpha)\ \sigma\cdot\tau\ d\mu\quad\forall\ \ \sigma,\tau\in\mathbf{L}^2(\Omega)$, then

$$\|\bar{a}\|_{\mathscr{L}(\mathbf{L}^2(\Omega),\mathbf{L}^2(\Omega);\mathbb{R})}\ \leqslant\ \|\alpha\|_{\mathrm{W}^{1,\infty}(\Omega)}\ , \tag{22a}$$

$$\bar{a}(\tau,\tau) \geqslant A\,\|\tau\|^2_{\mathbf{L}^2(\Omega)} \quad \forall\ \tau \in \mathbf{L}^2(\Omega)\,. \tag{22b}$$

Proof.
It is easy to see, that

$$P[\Omega_k](\alpha) = \frac{\int\limits_{\Omega_k} \alpha\, d\mu}{\mu(\Omega_k)} \leqslant \|\alpha\|_{\mathrm{L}^\infty(\Omega_k)}\,,$$

this implies (22a). Inequality (22b) follows from

$$P[\Omega_k](\alpha) = \frac{\int\limits_{\Omega_k} \alpha\, d\mu}{\mu(\Omega_k)} \geqslant A\,.\ \square$$

5.4. The properties of $\bar{a}_q$.

We discuss the interaction between Π, $\tilde{\Pi}$ and $\bar{a}_q$. We show, that $\bar{a}_q$ is $\mathbf{L}^2(\Omega)$-bounded on V_h. We also show, that $\bar{a}_q$ is $\mathbf{L}^2(\Omega)$-elliptic on V_h and X_h. We first notice, that the definitions of Π_h, $\tilde{\Pi}_h$ and $\bar{a}_q$ imply:

$$\bar{a}_q(\Pi_h\sigma,\Pi_h\tau) = \bar{a}_q(\sigma,\Pi_h\tau) = \bar{a}_q(\Pi_h\sigma,\tau) = \bar{a}_q(\sigma,\tilde{\Pi}_h\tau) = \bar{a}_q(\tilde{\Pi}_h\sigma,\tau) = \bar{a}_q(\tilde{\Pi}_h\sigma,\tilde{\Pi}_h\tau)\,. \tag{23}$$

Lemma 5.
If $\sigma,\tau \in \Sigma$, then

$$\bar{a}_q(\Pi_h\sigma,\Pi_h\sigma) \geqslant A\,\|\Pi_h\sigma\|^2_{\mathbf{L}^2(\Omega)}\,, \tag{24a}$$

$$\bar{a}_q(\tilde{\Pi}_h\sigma,\tilde{\Pi}_h\sigma) \geqslant \frac{A}{4}\,\|\tilde{\Pi}_h\sigma\|^2_{\mathbf{L}^2(\Omega)}\,, \tag{24b}$$

$$\bar{a}_q(\Pi_h\sigma,\Pi_h\tau) \leqslant 3\,\|\alpha\|_{\mathrm{L}^\infty(\Omega)}\,\|\Pi_h\sigma\|_{\mathbf{L}^2(\Omega)}\,\|\Pi_h\tau\|_{\mathbf{L}^2(\Omega)}\,, \tag{24c}$$

$$\frac{1}{4}\bar{a}(\tilde{\Pi}_h\tau,\tilde{\Pi}_h\tau) \leqslant \bar{a}(\Pi_h\tau,\Pi_h\tau) \leqslant \bar{a}_q(\Pi_h\tau,\Pi_h\tau)\,. \tag{24d}$$

Proof.
We define some auxiliary variables, $a_{k,i} = P[\Gamma_{k,i,0}](\sigma)$, $b_{k,i} = P[\Gamma_{k,i,1}](\sigma)$, $c_{k,i} = P[\Gamma_{k,i,0}](\tau)$ and $d_{k,i} = P[\Gamma_{k,i,1}](\tau)$. According to (23) and (17), we have

$$\bar{a}_q(\Pi_h\sigma,\Pi_h\sigma) = \bar{a}_q(\tilde{\Pi}_h\sigma,\tilde{\Pi}_h\sigma) = \sum_{k\in K} P[\Omega_k](\alpha)\frac{\mu(\Omega_k)}{2}\sum_{i=1}^{2}(a^2_{k,i} + b^2_{k,i})\,.$$

Now, equation (24a) follows from (20a) and (24b) follows from (20b).

We use lemma 2 and Cauchy-Schwartz twice to obtain

$$\bar{a}_q(\Pi\sigma,\Pi\tau) = \sum_{k\in K} P[\Omega_k](\alpha)\frac{\mu(\Omega_k)}{2}\sum_{i=1}^{2}(a_{k,i}c_{k,i} + b_{k,i}d_{k,i}) \leqslant$$

$$\sum_{k\in K} P[\Omega_k](\alpha)\frac{\mu(\Omega_k)}{2}\left[\sum_{i=1}^{2}(a^2_{k,i} + b^2_{k,i})\right]^{1/2}\left[\sum_{i=1}^{2}(c^2_{k,i} + d^2_{k,i})\right]^{1/2} \leqslant$$

$$3\,\|\alpha\|_{\mathrm{L}^\infty(\Omega)}\sum_{k\in K}\|\Pi_h\sigma\|_{\mathbf{L}^2(\Omega_k)}\,\|\Pi_h\tau\|_{\mathbf{L}^2(\Omega_k)} \leqslant 3\,\|\alpha\|_{\mathrm{L}^\infty(\Omega)}\,\|\Pi_h\sigma\|_{\mathbf{L}^2(\Omega)}\,\|\Pi_h\tau\|_{\mathbf{L}^2(\Omega)}\,.$$

This proves (24c). Next, we verify (24d),

$$\bar{a}(\Pi_h\tau,\Pi_h\tau) = \sum_{k\in K} P[\Omega_k](\alpha)\mu(\Omega_k)\sum_{i=1}^{2}\left[\frac{c^2_{k,i}}{3} + \frac{d^2_{k,i}}{3} + \frac{c_{k,i}d_{k,i}}{6} + \frac{d_{k,i}c_{k,i}}{6}\right] \leqslant$$

$$\sum_{k\in K} P[\Omega_k](\alpha)\frac{\mu(\Omega_k)}{2}\sum_{i=1}^{2}(c^2_{k,i} + d^2_{k,i})\,,$$

and

$$\bar{a}(\tilde{\Pi}_h\tau,\tilde{\Pi}_h\tau) = \sum_{k\in K} P[\Omega_k](\alpha)\mu(\Omega_k)\sum_{i=1}^{2}\left[c_{k,i}^2(1-\eta_{k,i})^2 + d_{k,i}^2\eta_{k,i}^2 + 2c_{k,i}d_{k,i}(1-\eta_{k,i})\eta_{k,i}\right] \leqslant$$

$$2\sum_{k\in K} P[\Omega_k](\alpha)\mu(\Omega_k)\sum_{i=1}^{2}(c_{k,i}^2 + d_{k,i}^2)\ .$$

Together with

$$\bar{a}_q(\Pi\tau,\Pi\tau) = \sum_{k\in K} P[\Omega_k](\alpha)\frac{\mu(\Omega_k)}{2}\sum_{i=1}^{2}(c_{k,i}^2 + d_{k,i}^2)\ ,$$

this implies (24d). □

5.5. The difference between $\bar{a}$ and $\bar{a}_q$.

For our error estimates, we need an upper bound for the difference between the value of $a(\sigma_h,\tau)$ and that of $\bar{a}_q(\sigma_h,\tau)$ for $\sigma_h \in V_h$, $\tau \in \mathbf{H}^1(\Omega)$. Because

$$|a(\sigma_h,\tau) - \bar{a}(\sigma_h,\tau)| \leqslant 2h_{\max}\,\|\alpha\|_{W^{1,\infty}(\Omega)}\,\|\sigma_h\|_{\mathbf{L}^2(\Omega)}\,\|\tau\|_{\mathbf{L}^2(\Omega)}\ ,$$

an estimate for $|\bar{a}(\sigma_h,\tau) - \bar{a}_q(\sigma_h,\tau)|$ suffices. Such an estimate is derived in lemma 7. In lemma 6, we prove a general inequality, needed in the proof of the second lemma.

Lemma 6.
Let $f \in C([0,h_1])$ and $g \in C^1([0,h_1]\times[0,h_2])$ then

$$\left|\int_{x=0}^{h_1}\int_{y=0}^{h_2} f(x)\left[g(x,y) - \frac{1}{h_2}\int_{z=0}^{h_2} g(0,z)\,dz\right] dy\,dx\right| \leqslant \tag{25}$$

$$\int_{x=0}^{h_1}|f(x)|\,dx\int_{y=0}^{h_2}\int_{w=0}^{h_1}|\partial g(w,y)/\partial w|\,dw\,dy\ ,$$

Proof.

$$\left|\int_{x=0}^{h_1}\int_{y=0}^{h_2} f(x)\left[g(x,y) - \frac{1}{h_2}\int_{z=0}^{h_2} g(0,z)\,dz\right] dy\,dx\right| =$$

$$\left|\int_{x=0}^{h_1} f(x)\int_{y=0}^{h_2}\left[g(x,y) - \frac{1}{h_2}\int_{z=0}^{h_2} g(0,z)\,dz\right] dy\,dx\right| =$$

$$\left|\int_{x=0}^{h_1} f(x)\int_{y=0}^{h_2}\left[g(x,y) - g(0,y)\right] dy\,dx\right| = \left|\int_{x=0}^{h_1} f(x)\int_{y=0}^{h_2}\int_{z=0}^{x}\partial g(w,y)/\partial w\,dw\,dy\,dx\right| \leqslant$$

$$\int_{x=0}^{h_1}|f(x)|\int_{y=0}^{h_2}\int_{w=0}^{x}|\partial g(w,y)/\partial w|\,dw\,dy\,dx \leqslant \int_{x=0}^{h_1}|f(x)|\int_{y=0}^{h_2}\int_{w=0}^{h_1}|\partial g(w,y)/\partial w|\,dw\,dy\,dx\ ,$$

This implies (25). □

We are now ready to derive the desired estimate.

Lemma 7.
Let $\sigma_h \in V_h$ and $\tau = (\tau_1,\tau_2)^T \in \mathbf{H}^1(\Omega)$, then

$$|\bar{a}(\sigma_h,\tau) - \bar{a}_q(\sigma_h,\tau)| \leqslant 6\,\|\alpha\|_{L^\infty(\Omega)}\,h_{\max}\,\|\sigma_h\|_{\mathbf{L}^2(\Omega)}\,\|\tau\|_{\mathbf{H}^1(\Omega)}\ . \tag{26}$$

Proof.
To simplify our notation, we introduce $a_{k,i} = P[\Gamma_{k,i,0}](\sigma_h)$, $b_{k,i} = P[\Gamma_{k,i,1}](\sigma_h)$, $\tau_{k,i,0} = P[\Gamma_{k,i,0}](\tau_i)$

and $\tau_{k,i,1} = P[\Gamma_{k,i,1}](\tau_i)$. We prove the lemma for τ with $\tau_1, \tau_2 \in C^1(\overline{\Omega})$, and extend by density.

We consider the difference between the two forms on one subdomain Ω_k with $P[\Omega_k](\alpha) = 1$.

$$\Big| \int_{\Omega_k} \sigma_h \cdot \tau \, d\mu - ½\mu(\Omega_k) \sum_{i=1}^{2} \Big[P[\Gamma_{k,i,0}](\sigma_{h,i}\tau_i) + P[\Gamma_{k,i,1}](\sigma_{h,i}\tau_i) \Big] \Big| =$$

$$\Big| \int_{\Omega_k} \sum_{i=1}^{2} \Big[(1-\xi_{k,i})a_{k,i} - \xi_{k,i}b_{k,i} \Big] \tau_i \, d\mu - ½\mu(\Omega_k) \sum_{i=1}^{2} \Big[P[\Gamma_{k,i,0}](a_{k,i}\tau_i) + P[\Gamma_{k,i,1}](b_{k,i}\tau_i) \Big] \Big| =$$

$$\Big| \int_{\Omega_k} \sum_{i=1}^{2} \Big[(1-\xi_{k,i})a_{k,i}\tau_i - \xi_{k,i}b_{k,i}\tau_i - ½a_{k,i}\tau_{k,i,0} - ½b_{k,i}\tau_{k,i,1} \Big] \, d\mu \Big| =$$

$$\Big| \int_{\Omega_k} \sum_{i=1}^{2} \Big[(1-\xi_{k,i})a_{k,i}(\tau_i - \tau_{k,i,0}) - \xi_{k,i}b_{k,i}(\tau_i - \tau_{k,i,1}) + (1-\xi-½)a_{k,i}\tau_{k,i,0} + (\xi-½)b_{k,i}\tau_{k,i,1} \Big] \, d\mu \Big| \leqslant$$

$$\sum_{i=1}^{2} \Big[\Big| \int_{\Omega_k} (1-\xi_{k,i}) \, a_{k,i}(\tau_i - \tau_{k,i,0}) \, d\mu \Big| + \Big| \int_{\Omega_k} \xi_{k,i} \, b_{k,i}(\tau_i - \tau_{k,i,1}) \, d\mu \Big| \Big] .$$

As a result of the application of lemma 6 to each term in this sum, we find

$$\frac{1}{2}\sum_{i=1}^{2} \Big[|a_{k,i}| h_{k,i} \int_{\Omega_k} |\partial\tau_i / \partial x_i| \, d\mu + |b_{k,i}| h_{k,i} \int_{\Omega_k} |\partial\tau_i / \partial x_i| \, d\mu \Big] \leqslant$$

$$\frac{1}{2}\sum_{i=1}^{2} h_{k,i} \int_{\Omega_k} (|a_{k,i}| + |b_{k,i}|) \, |\partial\tau_i / \partial x_i| \, d\mu \leqslant$$

$$\frac{1}{2} h_{k,\max} \sum_{i=1}^{2} \| \, |a_{k,i}| + |b_{k,i}| \, \|_{L^2(\Omega_k)} \| \partial\tau_i / \partial x_i \|_{L^2(\Omega_k)} \leqslant$$

$$h_{k,\max} \sum_{i=1}^{2} \| (a_{k,i}^2 + b_{k,i}^2)^{½} \|_{L^2(\Omega_k)} \| \partial\tau_i / \partial x_i \|_{L^2(\Omega_k)} \leqslant$$

$$h_{k,\max} \| (\sum_{i=1}^{2} (a_{k,i}^2 + b_{k,i}^2))^{½} \|_{L^2(\Omega_k)} \| (\sum_{i=1}^{2} (\partial\tau_i / \partial x_i)^2)^{½} \|_{L^2(\Omega_k)} \leqslant$$

$$6 h_{k,\max} \| \sigma_h \|_{\mathbf{L}^2(\Omega_k)} \| \tau \|_{\mathbf{H}^1(\Omega_k)} \cdot$$

The application of the Cauchy-Schwartz inequality to this last term and insertion of α yields the following result,

$$| \overline{a}(\sigma_h, \tau) - \overline{a}_q(\sigma_h, \tau) | \leqslant 6 h_{\max} \| \alpha \|_{L^\infty(\Omega)} \| \sigma_h \|_{L^2(\Omega)} \| \tau \|_{H^1(\Omega)} \cdot$$

Because $C^1(\overline{\Omega})$ is dense in $\mathrm{H}^1(\Omega)$, the formula also holds for $\tau_1, \tau_2 \in \mathrm{H}^1(\Omega)$.

□

6 The error estimates.

We use the standard estimates for $\| \sigma - \Pi_h \sigma \|_{L^2(\Omega)}$ and $\| u - P_h u \|_{L^2(\Omega)}$, as described in section 3.4, to reduce the problem to deriving bounds for $\| P_h u - u_h \|_{L^2(\Omega)}$ and $\| \Pi_h \sigma - \sigma_h \|_{L^2(\Omega)}$. We discuss two possible derivations of an $\mathcal{O}(h)$ error bound. The first derivation needs the assumption, that $h_{\max}$ is "small enough", the second derivation places a condition on an approximation of the discrete version of the adjoint problem.

6.1. The first estimate.

The following two lemmas show special properties of our discretisation. We need these properties to derive the error bound.

Lemma 8.

Let $\tau \in \Sigma$, $t \in \mathrm{L}^2(\Omega)$, then

$$\overline{b}(\tilde{\Pi}_h\tau, t-P_h t) - (div\ \tilde{\Pi}_h\tau, t-P_h t) = 0 \ . \tag{27}$$

Proof.
This is true, because $\mathbf{P}_h(\vec{\beta})\cdot\tilde{\Pi}_h\tau \ - \ div\ \tilde{\Pi}_h\tau$ is constant on Ω_k. □

Lemma 9.
If (σ,u) is a solution of (6) and (σ_h,u_h) is a solution of (18), then

$$(div\ (\sigma-\sigma_h), P_h t) + c(u-u_h, P_h t) = 0 \quad \forall \ t \in \mathrm{L}^2(\Omega) \ . \tag{28}$$

Proof.
We take (18b),

$$(div\ \sigma_h, P_h t) + \overline{c}(u_h, P_h t) = (f, P_h t) \ ,$$

$\overline{c}$ is derived by orthogonal $\mathrm{L}^2(\Omega_k)$ projection, so this implies

$$(div\ \sigma_h, P_h t) + c(u_h, P_h t) = (f, P_h t) \ .$$

If we subtract this from (6b), $(div\ \sigma, P_h t) + c(u, P_h t) = (f, P_h t)$, then we find (28). □

We are now ready to give an estimate for $\|\Pi_h\sigma-\sigma_h\|_{\mathrm{L}^2(\Omega)}$.

Theorem 1.
If (σ,u) is the solution of (6), (σ_h,u_h) is the solution of (18) and $(\sigma,u) \in \mathbf{H}^1(\Omega)\times \mathrm{H}^2(\Omega)$, then

$$\exists \ C_h > 0 \, , D > 0 : \ \|\Pi_h\sigma - \sigma_h\|_{\mathrm{L}^2(\Omega)} \leqslant C_h + D\,\|Pu - u_h\|_{\mathrm{L}^2(\Omega)} \ , \tag{29}$$

where C_h is $\mathcal{O}(h_{\max})$ and D is independent of $h_{\max}$.
Proof.
According to (24a), $A\,\|\Pi_h\sigma-\sigma_h\|^2_{\mathrm{L}^2(\Omega)} \leqslant \overline{a}_q(\Pi_h\sigma-\sigma_h,\ \Pi_h\sigma-\sigma_h)$. From the definition of $\overline{a}_q$, Π_h and $\tilde{\Pi}_h$, it follows, that

$$\overline{a}_q(\Pi_h\sigma-\sigma_h, \Pi_h\sigma-\sigma_h) = \overline{a}_q(\Pi_h\sigma-\sigma_h, \tilde{\Pi}_h(\sigma-\sigma_h)) = \overline{a}_q(\sigma-\sigma_h, \tilde{\Pi}_h(\sigma-\sigma_h)) \ .$$

This is the starting point for the derivation of our error bound. Equations (6a) and (18a) imply, that

$$\begin{gathered}\overline{a}_q(\sigma-\sigma_h, \tilde{\Pi}_h(\sigma-\sigma_h)) = (\overline{a}_q-a)(\sigma, \tilde{\Pi}_h(\sigma-\sigma_h)) + a(\sigma, \tilde{\Pi}_h(\sigma-\sigma_h)) - \overline{a}_q(\sigma_h, \tilde{\Pi}_h(\sigma-\sigma_h)) = \\ (\overline{a}_q-a)(\sigma, \tilde{\Pi}_h(\sigma-\sigma_h)) + (div\ \tilde{\Pi}_h(\sigma-\sigma_h), u) - b(\tilde{\Pi}_h(\sigma-\sigma_h), u) + \\ < g, \vec{n}_{\partial\Omega}\cdot\tilde{\Pi}_h(\sigma-\sigma_h) > \ + \overline{b}(\tilde{\Pi}_h(\sigma-\sigma_h), u_h) - (div\ \tilde{\Pi}_h(\sigma-\sigma_h), u_h) - \ < g, \vec{n}_{\partial\Omega}\cdot\tilde{\Pi}_h(\sigma-\sigma_h) > \ = \\ (\overline{a}_q-a)(\sigma, \tilde{\Pi}_h(\sigma-\sigma_h)) + (div\ \tilde{\Pi}_h(\sigma-\sigma_h), u)_{\mathrm{L}^2(\Omega)} - (b-\overline{b})(\tilde{\Pi}_h(\sigma-\sigma_h), u) - \overline{b}(\tilde{\Pi}_h(\sigma-\sigma_h), u) + \\ \overline{b}(\tilde{\Pi}_h(\sigma-\sigma_h), u_h) - (div\ \tilde{\Pi}_h(\sigma-\sigma_h), u_h) \ .\end{gathered}$$

Where we give $b-\overline{b}$, $\overline{a}_q-a$ etc. their obvious meaning. If we use lemma 8, we find:

$$\begin{gathered}A\,\|\Pi_h\sigma-\sigma_h\|^2_{\mathrm{L}^2(\Omega)} \leqslant (\overline{a}_q-a)(\sigma, \tilde{\Pi}_h(\sigma-\sigma_h)) - (b-\overline{b})(\tilde{\Pi}_h(\sigma-\sigma_h), u) + \\ (div\ \tilde{\Pi}_h(\sigma-\sigma_h), P_h u-u_h)_{\mathrm{L}^2(\Omega)} - \overline{b}(\tilde{\Pi}_h(\sigma-\sigma_h), P_h u-u_h) \ .\end{gathered}$$

If we use (19b) and (19c) to prepare the way, then the application of lemma 9 to this expression results in:

$$\begin{gathered}A\,\|\Pi_h\sigma-\sigma_h\|^2_{\mathrm{L}^2(\Omega)} \leqslant (\overline{a}_q-a)(\sigma, \tilde{\Pi}_h(\sigma-\sigma_h)) - (b-\overline{b})(\tilde{\Pi}_h(\sigma-\sigma_h), u) - \\ c(u-u_h, P_h u-u_h) - \overline{b}(\tilde{\Pi}_h(\sigma-\sigma_h), Pu-u_h) \ .\end{gathered}$$

This can be written as:

$$\begin{gathered}A\,\|\Pi_h\sigma-\sigma_h\|^2_{\mathrm{L}^2(\Omega)} \leqslant \\ (\overline{a}_q-\overline{a})(\sigma, \Pi_h(\sigma-\sigma_h)) - \overline{a}(\sigma, (\tilde{\Pi}_h-\Pi_h)(\sigma-\sigma_h)) + (\overline{a}-a)(\sigma, \tilde{\Pi}_h(\sigma-\sigma_h)) - (b-\overline{b})(\tilde{\Pi}_h(\sigma-\sigma_h), u) -\end{gathered}$$

$$(c-\overline{c})(u-P_h u,P_h u-u_h) \; - \; c(P_h u-u_h,P_h u-u_h) \; - \; \overline{b}(\Pi_h(\sigma-\sigma_h),P_h u-u_h)\,.$$

Using the estimates from lemma 7, lemma 4 and lemma 3, we obtain:

$$A\,\|\Pi_h\sigma - \sigma_h\|^2_{L^2(\Omega)} \leqslant$$
$$6h_{\max}\|\alpha\|_{L^\infty(\Omega)}\|\sigma\|_{H^1(\Omega)}\|\Pi_h(\sigma-\sigma_h)\|_{L^2(\Omega)} + \|\alpha\|_{L^\infty(\Omega)}\|\sigma\|_{L^2(\Omega)}5h_{\max}\|\vec{\beta}\|_{L^\infty(\Omega)}\|\Pi_h\sigma-\sigma_h\|_{L^2(\Omega)} +$$
$$2h_{\max}\|\alpha\|_{W^{1,\infty}(\Omega)}\|\sigma\|_{L^2(\Omega)}\|\tilde{\Pi}_h(\sigma-\sigma_h)\|_{L^2(\Omega)} + 4h_{\max}\|\vec{\beta}\|_{W^{1,\infty}(\Omega)}\|\tilde{\Pi}_h(\sigma - \sigma_h)\|_{L^2(\Omega)}\|u\|_{L^2(\Omega)} +$$
$$\|\vec{\beta}\|_{L^\infty(\Omega)}\|\Pi_h\sigma - \sigma_h\|_{L^2(\Omega)}\|P_h u-u_h\|_{L^2(\Omega)} +$$
$$\|\gamma\|_{W^{1,\infty}(\Omega)}\,(2h_{\max}\|u-P_h u\|_{L^2(\Omega)} + \|P_h u-u_h\|_{L^2(\Omega)})\,\|P_h u-u_h\|_{L^2(\Omega)}\,,$$

where we used, that:

$$\|w-P_h(w)\|_{L^\infty(\Omega)} \leqslant 2h_{\max}\|w\|_{W^{1,\infty}(\Omega)} \quad \forall \; w \in W^{1,\infty}(\Omega)\,.$$

We use lemma 2 to replace $\tilde{\Pi}_h$ by Π_h,

$$A\,\|\Pi_h\sigma - \sigma_h\|^2_{L^2(\Omega)} \leqslant$$
$$\|\sigma\|_{H^1(\Omega)}\left[\|\alpha\|_{W^{1,\infty}(\Omega)}(30h_{\max} + 5h_{\max}\|\vec{\beta}\|_{L^\infty(\Omega)}) + 48h_{\max}\|\vec{\beta}\|_{W^{1,\infty}(\Omega)}\|u\|_{L^2(\Omega)}\right]\|\Pi_h(\sigma-\sigma_h)\|_{L^2(\Omega)} +$$
$$\|\vec{\beta}\|_{L^\infty(\Omega)}\|\Pi_h\sigma - \sigma_h\|_{L^2(\Omega)}\|P_h u-u_h\|_{L^2(\Omega)} +$$
$$\|\gamma\|_{W^{1,\infty}(\Omega)}\,(2h_{\max}\|u-P_h u\|_{L^2(\Omega)} + \|P_h u-u_h\|_{L^2(\Omega)})\,\|P_h u-u_h\|_{L^2(\Omega)}\,.$$

If we define

$$C_1 := (\|\sigma\|_{H^1(\Omega)} + \|u\|_{L^2(\Omega)})\left[\|\alpha\|_{W^{1,\infty}(\Omega)}(30 + 5\|\vec{\beta}\|_{L^\infty(\Omega)}) + 48\|\vec{\beta}\|_{W^{1,\infty}(\Omega)}\right],$$
$$C_2 := \max\left[1\,,\ \|\beta\|_{L^\infty(\Omega)}\,,\ \|\gamma\|_{L^\infty(\Omega)}\,,\ \|\gamma\|_{L^\infty(\Omega)}\|u-P_h u\|_{L^2(\Omega)}\right],$$

then the above expression can be written as:

$$A\,\|\Pi_h\sigma - \sigma_h\|^2_{L^2(\Omega)} \leqslant h_{\max}C_1\|\Pi_h(\sigma-\sigma_h)\|_{L^2(\Omega)} +$$
$$C_2(\|\Pi_h\sigma - \sigma_h\|_{L^2(\Omega)} + h_{\max} + \|P_h u - u_h\|_{L^2(\Omega)})\|P_h u-u_h\|_{L^2(\Omega)}\,.$$

If we use the following notation,

$$x := \|P_h u-u_h\|_{L^2(\Omega)}\,,\, y := \|\Pi_h\sigma-\sigma_h\|_{L^2(\Omega)}\,,$$

then we find

$$Ay^2 \leqslant C_1 h_{\max} y + C_2(y + h_{\max} + x)x \leqslant$$
$$C_2(C_1 h_{\max} + x)y + C_2(h_{\max} + x)x \leqslant C_2(C_1 h_{\max} + x)y + C_2(h_{\max} + x)(h_{\max} + x)\,,$$

This implies, that

$$\frac{A}{C_2}y^2 - y(C_1 h_{\max} + x) - (h_{\max} + x)^2 \leqslant 0\,.$$

We know, that A , C_2 , x , $y \in]0,\infty[$, so y must lie between 0 and the positive root of this polynomial,

$$0 < y < \frac{C_2}{2A}((C_1 h_{\max} + x) + ((C_1 h_{\max} + x)^2 + 4(h_{\max} + x)^2)^{1/2}) \leqslant \frac{C_2}{2A}((2+C_1)h_{\max} + 3x)\,.$$

From this, (29) follows immediately. □

Next, we prepare for the second part of our error estimate.

Lemma 10.
If (σ,u) is the solution of (6), (σ_h,u_h) is a solution of (18) and (τ,q) is the solution of the adjoint problem for an arbitrary right hand side $p \in L^2(\Omega)$, then

$$(div\ \tilde{\Pi}_h\tau, P_h u - u_h) - \bar{b}(\tilde{\Pi}_h\tau, P_h u - u_h) + c(u - u_h, P_h q) =$$
$$a(\sigma, \tilde{\Pi}_h\tau - \tau) + a(\sigma - \Pi_h\sigma, \tau) + (\bar{a}_q - a)(\sigma - \sigma_h, \tau) - (\bar{a}_q - a)(\sigma, \tau) + (b - \bar{b})(\tilde{\Pi}_h\tau, u)\ .$$

Proof.
According to lemma 8 and lemma 9,

$$(div\ \tilde{\Pi}_h\tau, P_h u - u_h) - \bar{b}(\tilde{\Pi}_h\tau, P_h u - u_h) + c(u - u_h, P_h q) =$$
$$(div\ \tilde{\Pi}_h\tau, u - u_h) - \bar{b}(\tilde{\Pi}_h\tau, u - u_h) - (div\ (\Pi_h\sigma - \sigma_h), P_h q)\ .$$

We apply (10a) to write this in another form and we use equation (6a) and equation (18a)

$$(div\ \tilde{\Pi}_h\tau, u - u_h) - \bar{b}(\tilde{\Pi}_h\tau, u - u_h) - a(\tau, \Pi_h\sigma - \sigma_h) =$$
$$a(\sigma, \tilde{\Pi}_h\tau) - \ < g, \tilde{\Pi}_h\tau\cdot\vec{n}_{\partial\Omega} > \ - \bar{a}_q(\sigma_h, \tilde{\Pi}_h\tau) +$$
$$< g, \tilde{\Pi}_h\tau\cdot\vec{n}_{\partial\Omega} > \ - a(\Pi_h\sigma - \sigma_h, \tau) + (b - \bar{b})(\tilde{\Pi}_h\tau, u) =$$
$$a(\sigma, \tilde{\Pi}_h\tau - \tau) + a(\sigma - \Pi_h\sigma, \tau) + (a - \bar{a}_q)(\sigma_h, \tau) + (b - \bar{b})(\tilde{\Pi}_h\tau, u) =$$
$$a(\sigma, \tilde{\Pi}_h\tau - \tau) + a(\sigma - \Pi_h\sigma, \tau) + (\bar{a}_q - a)(\sigma - \sigma_h, \tau) - (\bar{a}_q - a)(\sigma, \tau) + (b - \bar{b})(\tilde{\Pi}_h\tau, u)\ .\ \ \square$$

Theorem 2.
If the adjoint problem (10) is regular, (σ, u) is the solution of (6), $(\sigma, u) \in \mathbf{H}^1(\Omega) \times H^2(\Omega)$, (σ_h, u_h) is a solution of (18) and $h_{\max}$ is small enough for the application of the results from section 3.4, then

$$\|P_h u - u_h\|_{L^2(\Omega)} \leqslant$$
$$h_{\max} C \Big[\|\alpha\|_{W^{1,\infty}(\Omega)} (\|\sigma\|_{L^2(\Omega)} + \|\sigma_h\|_{L^2(\Omega)}) +$$
$$(\|\vec{\beta}\|_{W^{1,\infty}(\Omega)} + \|\gamma\|_{L^\infty(\Omega)}) \|P_h u - u_h\|_{L^2(\Omega)} \Big]\ .$$

Proof.
If we have an estimate for $(P_h u - u_h, p)$ for all $p \in L^2(\Omega)$, then we can use

$$\|t\|_{L^2(\Omega)} = \sup_{p \in L^2(\Omega), p \neq 0} \frac{(p, t)}{\|p\|_{L^2(\Omega)}}\ ,$$

to find $\|P_h u - u_h\|_{L^2(\Omega)}$. We use the regularity of the adjoint problem (10) to find $(\tau, w) \in \mathbf{H}^1(\Omega) \times L^2(\Omega)$, such that

$$(p, P_h u - u_h) = (div\ \tau, P_h u - u_h) - b(\tau, P_h u - u_h) + c(w, P_h u - u_h)\ .$$

If we rearrange the terms and apply lemma 10, we obtain

$$(p, P_h u - u_h) = (div\ \tilde{\Pi}_h\tau, P_h u - u_h) - \bar{b}(\tilde{\Pi}_h\tau, P_h u - u_h) + \bar{c}(P_h w, u - u_h) +$$
$$(\bar{b} - b)(\tilde{\Pi}_h\tau, P_h u - u_h) - b(\tau - \tilde{\Pi}_h\tau, P_h u - u_h) + c(w - P_h w, P_h u - u_h) =$$
$$a(\sigma, \tilde{\Pi}_h\tau - \tau) + a(\sigma - \Pi_h\sigma, \tau) - (\bar{a}_q - a)(\sigma_h, \tau) -$$
$$(\bar{b} - b)(\tilde{\Pi}_h\tau, u) + (\bar{b} - b)(\tilde{\Pi}_h\tau, P_h u - u_h) - b(\tau - \tilde{\Pi}_h\tau, P_h u - u_h) + c(w - P_h w, P_h u - u_h)\ .$$

We can use the regularity of the adjoint problem (10), lemma 7 and the projection error estimates (15) and (16), to obtain

$$\sup_{p \in L^2(\Omega), p \neq 0} \frac{(p, P_h u - u_h)}{\|p\|_{L^2(\Omega)}} =$$
$$\|P_h u - u_h\|_{L^2(\Omega)} \leqslant$$
$$C h_{\max} \Big[\|\alpha\|_{W^{1,\infty}(\Omega)} \|\sigma\|_{\mathbf{H}^1(\Omega)} + \|\alpha\|_{W^{1,\infty}(\Omega)} \|\sigma_h\|_{L^2(\Omega)} +$$
$$\|\vec{\beta}\|_{W^{1,\infty}(\Omega)} \|u\|_{L^2(\Omega)} + \|\vec{\beta}\|_{W^{1,\infty}(\Omega)} \|P_h u - u_h\|_{L^2(\Omega)} + \|\gamma\|_{L^\infty(\Omega)} \|P_h u - u_h\|_{L^2(\Omega)} \Big]\ .\ \square$$

If $h_{\max}$ is small enough, theorem 1 and theorem 2 together give an $\mathcal{O}(h)$ error estimate. The use of the results from section 3.4 on the solution of the adjoint problem may limit the range of allowable $h_{\max}$. A more important limit on $h_{\max}$ is implied by the form of the estimate for $\| Pu - u_h \|_{L^2(\Omega)}$. The main problem is, that larger values of $\| \vec{\beta} \|_{W^{1,\infty}(\Omega)}$ decrease the range of $h_{\max}$ for which the estimate is valid. This problem can be avoided if we make an extra assumption. We discuss this in the next section.

6.2. A different approach.

To improve our estimate of $\| P_h u - u_h \|_{L^2(\Omega)}$, we consider the adjoint of the discrete problem. This means, that we look for $(\tau_h, v_h) \in X_h \times W_h$, such that

$$\bar{a}_q(\tau_h, \sigma) - (div\ \sigma, v_h) = 0 \quad \forall \ \ \sigma \in V_h \ , \tag{30a}$$

$$(div\ \tau_h, t) - \bar{b}(\tau_h, t) + \bar{c}(v_h, t) = (f, t) \quad \forall \ \ t \in W_h \ . \tag{30b}$$

We call this system regular, if there is at least one solution for each $f \in P_h(L^2(\Omega))$, and that all solutions for a particular f satisfy

$$\max(\| \Pi \tau_h \|_{L^2(\Omega)} \ , \ \| \tau_h \|_{L^2(\Omega)}) + \| v_h \|_{L^2(\Omega)} \leqslant C \| P_h f \|_{L^2(\Omega)} \ , \tag{30c}$$

with C independent of the mesh size. This is a somewhat less stringent regularity condition than that given for the continuous adjoint problem (10). Note, that $\tau_h \in X_h$, so $\tau_{h,i}$ is an exponential function on Ω_k for $i = 1,2$.

An example of a general condition under which this system is regular is the following:

$$\alpha \geqslant A > 0 \, , \gamma \geqslant C_0 > 0 \ \text{ and } \ AC_0 - \| \vec{\beta} \|^2_{\mathbf{L}^\infty(\Omega)} \geqslant C_1 > 0 \ . \tag{31}$$

To show this, we need the following relations,

$$\int_\Omega \frac{P_h(\alpha)}{4} \tau_h \cdot \tau_h - \mathbf{P}_h(\vec{\beta}) \tau_h v_h + P_h(\gamma) v_h v_h \ d\mu = \tag{32}$$

$$\int_\Omega \frac{P_h(\alpha)}{4} \left[\tau_h - \frac{2\mathbf{P}_h(\vec{\beta})}{P_h(\alpha)} v_h \right]^2 + \left[P_h(\gamma) - \frac{\mathbf{P}_h(\vec{\beta})^2}{P_h(\alpha)} \right] v_h v_h \ d\mu \geqslant \tag{32a}$$

$$\int_\Omega P_h(\gamma) \left[v_h - \frac{\mathbf{P}_h(\vec{\beta}) \cdot \tau_h}{2 P_h(\gamma)} \right]^2 + \left[P_h(\alpha) - \frac{\mathbf{P}_h(\vec{\beta})^2}{P_h(\gamma)} \right] \frac{\tau_h \cdot \tau_h}{4} \ d\mu \ . \tag{32b}$$

We know, that $(div\ \tilde{\Pi}_h \sigma, P_h t) = (div\ \Pi_h \sigma, P_h t)$, so, if we sum of (30a) and (30b) with $\sigma = \Pi_h \tau_h$ and $t = v_h$, we find

$$\bar{a}_q(\tau_h, \Pi_h \tau_h) - \bar{b}(\tau_h, v_h) + \bar{c}(v_h, v_h) = (f, v_h) \ . \tag{33}$$

According to (23), $\bar{a}_q(\tau_h, \Pi_h \tau_h) = \bar{a}_q(\tilde{\Pi}_h \tau_h, \tilde{\Pi}_h \tau_h)$, and by (24d) we have

$$\frac{1}{4} \bar{a}(\tilde{\Pi}_h \sigma, \tilde{\Pi}_h \sigma) \leqslant \bar{a}_q(\tilde{\Pi}_h \sigma, \tilde{\Pi}_h \sigma) \ .$$

Hence

$$\int_\Omega \frac{P_h(\alpha)}{4} \tau_h \cdot \tau_h - \mathbf{P}_h(\vec{\beta}) \tau_h v_h + P_h(\gamma) v_h v_h \ d\mu \leqslant \int_\Omega P_h(f) v_h \ d\mu \ . \tag{34}$$

This expression is identical to (32), so (32a) is smaller than (f, v_h), combined with (31) this implies

$$\frac{C_1}{\| \alpha \|_{L^\infty(\Omega)}} \| v_h \|_{L^2(\Omega)} \leqslant \| f \|_{L^2(\Omega)} \ . \tag{35a}$$

In the same way, we find, that (32b) is smaller than (f, v_h), together with (31) and (35a), this implies

$$\frac{C_1}{(\| \alpha \|_{L^\infty(\Omega)} \| \gamma \|_{L^\infty(\Omega)})^{1/2}} \| \tau_h \|_{\mathbf{L}^2(\Omega)} \leqslant \| f \|_{L^2(\Omega)} \ . \tag{35b}$$

It now remains to check the bound on $\| \Pi_h \tau_h \|$. We consider the contribution of one subdomain Ω_k. If $\| \Pi \tau_h \|_{L^2(\Omega_k)} \leqslant \| \tau_h \|_{L^2(\Omega_k)}$, then (23) and (24d) imply,

$$\bar{a}_q(\chi_k \tau_h, \Pi \tau_h) - \bar{b}(\chi_h \tau_h, v_h) + \bar{c}(\chi_k v_h, v_h) \geqslant$$

$$\int_{\Omega_k} \frac{P_h(\alpha)}{4} \tau_h \cdot \tau_h - \mathbf{P}_h(\vec{\beta}) \cdot \tau_h v_h + P_h(\gamma) v_h v_h \, d\mu \geqslant$$

$$\int_{\Omega_k} \frac{P_h(\alpha)}{4} \tau_h \cdot \tau_h - \frac{(\mathbf{P}_h(\vec{\beta}) \cdot \tau_h)^2}{P_h(\gamma)} + P_h(\gamma) \left[\mathbf{P}_h(\vec{\beta}) \cdot \tau_h - v_h \right]^2 d\mu \geqslant$$

$$\int_{\Omega_k} \left[\frac{P_h(\alpha)}{4} - \frac{\mathbf{P}_h(\vec{\beta})^2}{P_h(\gamma)} \right] \tau_h \cdot \tau_h - + P_h(\gamma) \left[\mathbf{P}_h(\vec{\beta}) \cdot \tau_h - v_h \right]^2 d\mu \geqslant$$

$$\frac{C_1}{\| \gamma \|_{L^\infty(\Omega)}} \| \tau_h \|^2_{L^2(\Omega_k)} \geqslant \frac{C_1}{\| \gamma \|_{L^\infty(\Omega)}} \| \Pi_h \tau_h \|^2_{L^2(\Omega_k)} .$$

If $\| \Pi \tau_h \|_{L^2(\Omega_k)} > \| \tau_h \|_{L^2(\Omega_k)}$, then (23) and (24d) imply

$$\bar{a}_q(\chi_k \tau_h, \Pi \tau_h) - \bar{b}(\chi_h \tau_h, v_h) + \bar{c}(\chi_k v_h, v_h) \geqslant$$

$$\int_{\Omega_k} \frac{P_h(\alpha)}{4} \Pi \tau_h \cdot \Pi \tau_h - \mathbf{P}_h(\vec{\beta}) \cdot \tau_h v_h + P_h(\gamma) v_h v_h \, d\mu \geqslant$$

$$\int_{\Omega_k} \frac{P_h(\alpha)}{4} \Pi_h \tau_h \cdot \Pi_h \tau_h - \frac{(\mathbf{P}_h(\vec{\beta}) \cdot \tau_h)^2}{P_h(\gamma)} + P_h(\gamma) \left[\mathbf{P}_h(\vec{\beta}) \cdot \tau_h + v_h \right]^2 d\mu \geqslant$$

$$\int_{\Omega_k} \left[\frac{P_h(\alpha)}{4} - \frac{\mathbf{P}_h(\vec{\beta})^2}{P_h(\gamma)} \right] \Pi_h \tau_h \cdot \Pi_h \tau_h - + P_h(\gamma) \left[\mathbf{P}_h(\vec{\beta}) \cdot \tau_h + v_h \right]^2 d\mu \geqslant \int_{\Omega_k} \frac{C_1}{\| \gamma \|_{L^\infty(\Omega)}} \| \Pi_h \tau_h \|_{L^2(\Omega_k)} .$$

We combine these two cases to find, that

$$\frac{C_1}{\| \gamma \|_{L^\infty(\Omega)}} \sum_{k \in K} \| \Pi_h \tau_h \|^2_{L^2(\Omega_k)} \leqslant \| f \|_{L^2(\Omega)} \| v_h \|_{L^2(\Omega)} .$$

We combine this with (35a) to obtain,

$$\frac{C_1}{(\| \alpha \|_{L^\infty(\Omega)} \| \gamma \|_{L^\infty(\Omega)})^{1/2}} \| \tau_h \|_{L^2(\Omega)} \leqslant \| f \|_{L^2(\Omega)} . \qquad (35c)$$

Theorem 3.
If we assume, that (30) is regular, then

$$\| P_h u - u_h \|_{L^2(\Omega)} \leqslant \qquad (36)$$

$$C \left[h_{\max} \| \alpha \|_{W^{1,\infty}(\Omega)} \| \sigma \|_{H^1(\Omega)} + h_{\max} \| \alpha \|_{L^\infty(\Omega)} \| \sigma \|_{H^1(\Omega)} 5 \| \vec{\beta} \|_{L^\infty(\Omega)} + \right.$$

$$\left. h_{\max} \| \vec{\beta} \|_{W^{1,\infty}(\Omega)} \| u \|_{L^2(\Omega)} + h_{\max} \| \gamma \|_{L^\infty(\Omega)} \| u \|_{L^2(\Omega)} \right] .$$

Proof.
We use regularity of (30) and (30b),

$$(P_h u - u_h, P_h f) = (div\ \tau_h, P_h u - u_h) - \bar{b}(\tau_h, P_h u - u_h) + \bar{c}(P_h u - u_h, v_h) .$$

Hence, according to lemma 8 and the definition of $\bar{c}$,

$$(P_h u - u_h, P_h f) = (div\ \tau_h, u - u_h) - \bar{b}(\tau_h, u - u_h) + \bar{c}(u - u_h, v_h) .$$

We use (6a) and (18a) to find

$$(P_h u - u_h, P_h f) = (div\ \tau_h, u - u_h) - (\bar{b} - b)(\tau_h, u) - b(\tau_h, u) + \bar{b}(\tau_h, u_h) + \bar{c}(u - u_h, v_h) =$$

$$(\bar{b}-b)(\tau_h,u) + a(\sigma,\tau_h) - \bar{a}_q(\sigma_h,\tau_h) + (\bar{c}-c)(u,v_h) + c(u - u_h,v_h) .$$

According to (23) and lemma 9, this implies

$$(P_h u - u_h,P_h f) = (\bar{b}-b)(\tau_h,u) + (a-\bar{a}_q)(\sigma,\tau_h) + \bar{a}_q(\Pi_h\sigma - \sigma_h,\tau_h) + (\bar{c}-c)(u,v_h) + (div\ (\Pi_h\sigma - \sigma_h),v_h) .$$

Now, (30a) implies,

$$(P_h u - u_h,P_h f) = (\bar{b}-b)(\tau_h,u) + (a-\bar{a}_q)(\sigma,\tau_h) + (\bar{c}-c)(u,v_h) =$$
$$(\bar{b}-b)(\tau_h,u) + (a-\bar{a})(\sigma,\tau_h) + \bar{a}(\sigma,\tau_h-\Pi_h\tau_h) + (\bar{a}-\bar{a}_q)(\sigma,\Pi_h\tau_h) + (\bar{c}-c)(u,v_h) .$$

Finally, we use lemma 3, lemma 7 and (30c) to obtain our error estimate (36). □

7 The one dimensional case.

Although the theorem is especially interesting for the two dimensional case, we consider here a one-dimensional example, because for this case, we can easily compare the discrete solution to a known continuous solution. We consider the equation (cf. eq. (1))

$$(\epsilon(u' + u/\epsilon))' = 0 \quad \text{on} \quad \Omega =]0,1[, \tag{37}$$

with boundary conditions

$$u(0) = 0 \text{ and } u(1) = U ,$$

and $\epsilon > 0$. Special care needs to be taken when we discretise this equation for $\epsilon \ll 1$. The exact solution is known,

$$u(x) = U(1 - \exp(-x/\epsilon))/(1 - \exp(-1/\epsilon)) , \tag{38}$$

$$\epsilon u'(x) + u(x) = U/(1-\exp(-1/\epsilon) . \tag{39}$$

Any solution of (37) generates a solution $(\sigma,u) \in V \times W$ for

$$\frac{1}{\epsilon}(\sigma,\tau) - (u,\tau') + \frac{1}{\epsilon}(u,\tau) = -\tau(1)U \quad \forall \ \tau \in V , \tag{40a}$$

$$(\sigma',t) = 0 \quad \forall \ t \in W . \tag{40b}$$

With $\sigma = -(\epsilon u' + u)$.

7.1. The discretisation.

We partition Ω into $N > 1$ intervals $\Omega_k =]x_{k-1},x_k[$ with $x_0 = 0$ and $x_N = 1$. We define $h_k = x_k - x_{k-1}$.

The functions η and ξ, introduced in section 3.3, become

$$\xi_k = \frac{x-x_{k-1}}{h_k} \quad \text{on } \Omega_k ,$$

$$\eta_k = \frac{\exp((x-x_{k-1})/\epsilon)-1}{\exp(h_k/\epsilon)-1} \quad \text{on } \Omega_k .$$

In the one dimensional case, the definitions in section 3.3 result in

$$V_h := \{ \tau \in \mathrm{H}^1(\Omega) \mid \tau = (1-\xi_k)a_k + \xi_k b_k \ \text{ on } \Omega_k \} , \tag{41}$$

$$W_h := \{ t \in W \mid t \text{ is constant on } \Omega_k \} , \tag{42}$$

$$X_h := \{ \tau \in \mathrm{H}^1(\Omega) \mid \tau = (1-\eta_k)a_k + \eta_k b_k \ \text{ on } \Omega_k \} . \tag{43}$$

Next, we have

$$(\sigma,\tau)_k = \int_{\Omega_k} \sigma\tau \, d\mu ,$$

and with quadrature, we obtain

$$(\sigma,\tau)_{k,q} = \frac{h_k}{2}\left[\sigma(x_{k-1})\tau(x_{k-1}) + \sigma(x_k)\tau(x_k)\right],$$

$$(\sigma,\tau)_q = \sum_{k\in K}(\sigma,\tau)_{q,k}.$$

Our discrete problem has the form:

$$\frac{1}{\epsilon}(\sigma_h,\tau)_q - (u_h,\tau') + \frac{1}{\epsilon}(u_h,\tau) = -\tau(1)U \quad \forall\ \tau\in X_h, \tag{44a}$$

$$(\sigma_h{}',t) = 0 \quad \forall\ t\in W_h. \tag{44b}$$

with $(\sigma_h,u_h)\in V_h\times W_h$. Equation (44b) implies, that σ_h is constant, we eliminate this equation and set $\sigma_h = \bar\sigma$.

$$\sum_{k=1}^{N}\left[\bar\sigma(1,\tau)_{q,k} + u_{h,k}\left[(1,\tau)_k - \epsilon(1,\tau')_k\right]\right] = -\epsilon\tau(1)U \quad \forall\ \tau\in V_h. \tag{45}$$

Explicitly, the discrete equations read,

$$\bar\sigma(1,1-\eta_1)_{q,1} + u_{h,1}\left[(1,1-\eta_1)_1 - \epsilon(1,-\eta_1{}')_1\right] = 0,$$

$$\bar\sigma(1,\eta_k)_{q,k} + \bar\sigma(1,1-\eta_{k+1})_{q,k+1} + u_{h,k}\left[(1,\eta_k)_k - \epsilon(1,\eta_k{}')_k\right] +$$
$$u_{h,k+1}\left[(1,1-\eta_{k+1})_{k+1} - \epsilon(1,-\eta_{k+1}{}')_{k+1}\right] = 0,$$
$$\text{for } k = 1,2,\ldots,N-1,$$

$$\bar\sigma(1,\eta_N)_N + u_{h,N}\left[(1,\eta_N)_N - \epsilon(1,\eta_N{}')_N\right] = -\epsilon U.$$

Assume $h_k = h$ for all intervals. We find,

$$\bar\sigma/2 + u_{h,1}\frac{\exp(h/\epsilon)}{\exp(h/\epsilon)-1} = 0,$$

$$\bar\sigma + u_{h,k}\frac{-1}{\exp(h/\epsilon)-1} + u_{h,k+1}\frac{\exp(h/\epsilon)}{\exp(h/\epsilon)-1} = 0,$$
$$\text{for } k = 1,2,\ldots,N-1,$$

$$\bar\sigma/2 + u_{h,N}\frac{-1}{\exp(h/\epsilon)-1} = -\frac{\epsilon}{h}U.$$

From the equations for two adjacent internal intervals we can eliminate $\bar\sigma$, we find:

$$u_{h,k}\frac{-1}{\exp(h/\epsilon)-1} + u_{h,k+1}\frac{\exp(h/\epsilon)}{\exp(h/\epsilon)-1} - u_{h,k-1}\frac{-1}{\exp(h/\epsilon)-1} - u_{h,k}\frac{\exp(h/\epsilon)}{\exp(h/\epsilon)-1} = 0,$$
$$\text{for } k = 2,3,\ldots,N-2.$$

This is equivalent with the Il'in [15] scheme.

$$\operatorname{cotanh}\left[\frac{h}{2\epsilon}\right](u_{h,k+1} - 2u_{h,k} + u_{h,k-1}) + (u_{h,k+1} - u_{h,k-1}) = 0 \text{ for } k = 2,3,\ldots,N-2. \tag{46}$$

The solution of the implied recursion relation is,

$$u_{h,k} = -U\frac{\epsilon}{h}\frac{\exp(h/\epsilon)-1}{1-\exp(-1/\epsilon)}\exp(-kh/\epsilon) - \bar\sigma,$$

with

$$\overline{\sigma} = \frac{1-\exp(h/\epsilon)}{1+\exp(h/\epsilon)}\frac{2\epsilon}{h}\;\frac{U}{1-\exp(-1/\epsilon)} = -\frac{2\epsilon}{h}\tanh\left[\frac{h}{2\epsilon}\right]\frac{U}{1-\exp(-1/\epsilon)} .$$

To see the effect of the quadrature rule, we compare this with $P_h u$,

$$P[\Omega_k](u) = -U\frac{\epsilon}{h}\frac{\exp(h/\epsilon)-1}{1-\exp(-1/\epsilon)}\exp(-kh/\epsilon) - \sigma$$

where

$$\sigma = -\frac{U}{1-\exp(-1/\epsilon)} .$$

We see, that the method is a cell centered version of the classical exponential fitting scheme. In this case the method is 2^{nd} order in h, but clearly the estimate is not uniform in ϵ. The error constant degenerates for $\epsilon \to 0$. However, the matrix for u_h, after elimination of σ_h through static condensation, is an M-matrix. Exact integration of $(1,1-\eta_1)_1$ and $(1,\eta_N)$ would have resulted in $\overline{\sigma} = \sigma$.

8 Conclusions.

The method has several good properties. For instance, just as for a finite volume method, if the true solution σ is divergence-free, then the same holds for σ_h. Furthermore we have an a priori error estimate, that depends linearly on $\|\vec{\beta}\|_{L^\infty(\Omega)}$. This in contrast to the approach, that uses Slotboom variables to obtain a symmetrical form of the equations. There we get a priori error estimates that depend on $\exp(\psi)$, where $\vec{\beta} = grad\,\psi$. (see, for example Brezzi[5]). In addition to this, after elimination of σ_h through static condensation the two dimensional discretisation results in an M-matrix for u_h. We can extend the method to three dimensions without additional difficulties.

References

1. D. L. Scharfetter and H. K. Gummel, "Large-Signal Analysis of a Silicon Read Diode Oscillator," *IEEE Transactions on Electron Devices* **ED-16**(1), pp. 64-77 (1969).
2. Siegfried Selberherr, *Analysis and simulation of semiconductor devices,* Springer-verlag, Wien New York (1984).
3. Peter A. Markowich, *The Stationary Semiconductor Device Equations,* Springer-Verlag, Wien New York (1986).
4. Wolfgang Fichtner, Donald J. Rose, and Randolph E. Bank, "Numerical method for semiconductor device simulation," *SIAM journal of scientific and statistical computing* **4**(3), pp. 416-435 (1983).
5. F. Brezzi, L. D. Marini, and P Pietra, "Two-dimensional exponential fitting and applicatins to drift-diffusion problems," *SIAM J. Numer. Anal.* **26**(6), pp. 1342-1355 (1989).
6. P. W. Hemker, *A Numerical Study of Stiff Two-Point Boundary Problems,* Mathematical Centre, Amsterdam (1977).
7. J. Douglas, Jr. and J. E. Roberts, "Global estimates for mixed methods for second order elliptic equations," *Mathematics of computation* **44**(169), pp. 39-52 (1985).
8. Robert A. Adams, *Sobolev Spaces,* Academic Press (1975).
9. V. Girault and P. Raviart, *Finite Element Methods for Navier-Stokes Equations,* Springer-Verlag (1986).
10. S. J. Polak, C. den Heijer, H. A. Schilders, and P. Markowich, "Semiconductor device modelling from the numerical point of view," *International Journal for Numerical Methods in Engineering* **24**, pp. 763-838 (1987).
11. Jean E. Roberts and Jean-Marie Thomas, "Mixed and Hybrid Finite Element Methods," RR 737, INRIA, Rocquencourt (October 1987).
12. P. A. Raviart and J. M. Thomas, "A mixed finite element method for 2-nd order elliptic

problems," in *Mathematical aspects of the finite element method*, Springer (1977).

13. J. C. Nedelec, "Mixed Finite Elements in $\mathbb{R}^3$," *Numer. Math.* **35**, pp. 315-341 (1980).

14. P. G. Ciarlet and P. A. Raviart, "General Lagrange and Hermite Interpolation in $\mathbb{R}^n$ with Applications to Finite Element Methods.," *Arch. Rational Mech. Anal.* **46**, pp. 177-199 (1972).

15. A. M. Il'in, "Differencing scheme for a differential equation with a small parameter affecting the highest derivative," *Mathematical Notes of the academy of sciences of the USSR* **6**(1-2), pp. 596-602 (1969).